Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# Characterization of the Distribution of Participation in Wikis

**ANTONIO TENORIO[1,2], JAVIER ARROYO[1,3], SAMER HASSAN[1,3,4],**

[1]Dpto Ingeniería del Software e Inteligencia Artificial, Facultad de Informática. Universidad Complutense de Madrid Spain
[2]Decentralized Science. Madrid, Spain
[3]Instituto de Tecnología del Conocimiento. Universidad Complutense de Madrid, Spain
[4]Berkman Klein Center for Internet and Society. Harvard University, USA

Corresponding author: Antonio Tenorio (e-mail: antoniotenorio@ucm.es).

## ABSTRACT

The distribution of participation in peer production, such as wikis and open source communities, has been traditionally characterized as a power-law distribution. This assumption has several implications, such as the assumed uniformity of the relation across participants, or the characterization of occasional vs core contributors. However, recent statistical studies on empirical data have challenged the power-law dominance in different domains. This work critically examines the assumption that the distribution of participation in wikis follows such distribution. We use statistical tools to exhaustively explore over 6,000 wikis from Wikia, the largest wiki repository. We analyse the empirical distribution of each wiki comparing it with different well-known skewed distributions. The results show that the power-law performs sensibly poor, surpassed by three others, while the truncated power-law is superior to all others or superior to some and as good as the rest in the 99.3% of the cases. In conclusion, these results refute the widely accepted assumption of power-law ubiquity in peer production, and aim to open a discussion around the issue.

## INDEX TERMS

## I. INTRODUCTION

Since the emergence of online communities, one of the major topics of interest is to understand the different levels in which members participate: that is, the distribution of participation, also named distribution of work, or effort. Far from classical organizational structures, and more similar to volunteer-driven social movements, communities show an inherent participation inequality across its participants. Specifically in peer production communities, such as those in wikis and free/open source software, this issue has derived multiple research questions: the concentration of participation in an elite [1]–[3], the degree of participation inequality [4]–[6], the characterization of who participates more [7], [8], the process of changing user roles [9], [10], or the evolution of participation depending on multiple factors [11], [12].

An important bulk of peer production research tends to say that the distribution of participation follows a power-law. Intuitively, this means a very small number of contributors would concentrate most of the participation (or work), highlighting participation inequality. Formally, a power law is a simple relationship between two quantities such that one is proportional to a fixed power of the other. In the issue at hand, i.e. participation, the two quantified dimensions are the number of contributions, and the share of people in the community that has made such number of contributions. The relationship among them is negative, that is, the higher the number of contributions, the smaller the share of contributors that has made such number of contributions. According to this idea, a small amount of contributions would be common, while larger amounts would be more rare. This fits with the assumption of participation inequality in which most members of the community tend to participate very little (occasional contributors), while a few of them account for an enormous amount of contributions (core contributors). In fact, the statement is not ungrounded, since several statistical studies focused on Wikipedia claim that the plot of edits per user follow a power law distribution [2], [13], and other studies find similar behavior in free/open source communities [14]–[17] or other peer production communities [18], [19].[1]

However, the power law implies an underlying regularity in the behavior of the phenomenon under study. In particular, the power relationship should hold independently of which

[1]Other studies just mention a highly skewed distribution or similar statements without further specification [20]–[22].

particular scale we are looking at. This may not be the case in real data. In fact, recent studies in statistics challenge the apparent dominance of power law across multiple fields with the help of modern sophisticated statistical tools [23], [24]. According to these works, power law distributions are complicated to detect because fluctuations occur in the tail of the distribution, and because of the difficulty of identifying the range over which power-law behavior holds.

In the peer production field, the regularity of the power law would imply that the relationship that holds for the occasional contributors would be the same to that for the core members, which may be a strong assumption for a community.

In particular, the tail of the distribution, which represents the activity of core contributors, may not have an extreme behaviour as the power law suggests, i.e., the number of extremely active contributors may not be as high. If that is the case, more conservative distributions, such as the the truncated power law, which was found suitable in a previous examination of wiki data [25], would provide a better fit.

According to these premises, it seems reasonable to question the characterization of the participation in peer production as a power law, and consider other heavy-tailed distributions. Thus, we will apply the statistical tools proposed in [24] to study peer production distributions, and more precisely participation distributions from wiki communities. The statistical tools proposed in that work provide a test to determine whether a distribution provides a better fit than another with respect to the empirical data provided. Thus, we will use them to analyze whether one candidate distribution consistently provides a better fit than the others. The candidates will be five well-known distributions, namely, the power law, three heavy-tailed distributions with a tail more conservative than the power law (truncated power law, stretched exponential and log-normal) and a non-heavy tailed distribution (exponential).

In our work, we focus on Wikia, the largest wiki repository which provides a large and diverse sample of peer production communities. Wikia accounts for over 300,000 wikis. However, because of constraints of the statistical methods used, which require a certain minimum of observations, we will use for our analysis the ~6,000 wikis which have at least 100 users.

The rest of the article proceeds as follows. Section II details the process followed to perform the statistical analysis and for the data collection. Section III shares the results of the statistical study of user contributions, and discusses its results through the explanation of series of graphs. Afterwards, Section IV offers an analysis of the winning distribution, i.e. the truncated power law, and proposes an interpretation of its parameters and how they characterize the different wikis under study. The paper closes with some concluding remarks and future work in Section V.

## II. METHODOLOGY AND DATA COLLECTION

### A. METHODOLOGY

Following [23] and [24], our study is divided in two analyses. First, in order to assess if the power law distribution is a plausible model for the given empirical data, we use the authors' goodness of fit test. Then, we perform an exhaustive analysis in order to identify which distribution better describes each wiki within the data set. These two methods are explained in this section.

#### 1) Goodness of fit

Clauset et al [23] propose a statistical test in order to asses if a distribution plausibly follows a power law. First, the test fits the dataset to a power law distribution model , finding its slope, or $\alpha$ parameter, and the minimum value from which the power law behavior is observed, or $x_{min}$ parameter.

Afterwards, a set of comparable synthetic data-sets that follow the distribution (i.e. have the same parameters) is created. The distance of the real data to its power law model is compared with the distance of the synthetic data sets to their power law models. Note that the synthetic datasets are also fit to power law models to compete in similar conditions . These distances are calculated using the Kolmogorov-Smirnov (KS) statistic. The goodness-of-fit test returns a p-value between 0 and 1 representing the number of synthetic dataset fits that outperformed the real data fit. E.g. a p-value of 0.4 represents that the real data fits better the power law than the 40% of the synthetically generated data. This p-value is then used to decide whether to rule out the hypothesis of the data following a power law. In our study, we rule out the power law model hypothesis if the p-value is smaller than 0.1, as [23] and [24] do, i.e. if the probability of obtaining a worse fit by chance is smaller than 10%. The number of synthetic data sets used to calculate the p-value determines the accuracy of the result. Following [23], for the result to be accurate to withing $\epsilon$, we should generate about $\epsilon^{-2}/4$ samples. Our study generates 100 synthetic data sets per test, therefore, the results are within an $\epsilon$ of 0.05.

When the number of observations is relatively small, this goodness of fit test cannot rule out a power law model in those cases in which the data follows other distributions such as the log-normal or exponential. For instance, for data following an exponential distribution with $\lambda = 0.125$, at least 100 observations are needed for the average p-value to drop bellow our threshold of 0.1, while for data following a log-normal distribution with $\mu = 0.3$, the average p-value drops below 0.1 from around 300 observations [23]. Thus, high p-values in these distributions with small number of observations should not be interpreted as the data following a power law. Moreover, as studied in the following section, even if a distribution plausibly follows a power law, other distributions may fit the data better. This work considers wikis with more than 100 observations (i.e. wikis with over 100 contributors) for the p-value study for two reasons. First, as already mentioned the goodness-of-fit test would not be able to rule out competing distributions. Second, as the wikis with less than 100 contributors represent more than 98% of

wikis (See Section II), the percentage of wikis passing the test due to the small number of observations may hinder the adequacy of the power-law hypothesis for those wikis with enough data to provide test results significant enough to distinguish from alternative models.

Summarizing, our study considers distributions with more than 100 observations (i.e. wikis with over 100 contributors), performs the goodness-of-fit tests proposed by [23] considering those with a p-value greater or equal to $0.1(\pm 0.0158)$ to plausibly follow a power law. The results of these tests are presented in Section III

This study was performed using the poweRlaw R package [26]. Besides, the R script source code, required for applying these statistical tests to our data, is available as free/open source software to facilitate replication[2].

#### 2) Likelihood-ratio test

The previously described goodness of fit test provides a tool to decide whether to rule out a power law distribution as a good model for the data. However, even if a power law model is not rejected, there may be better alternative distributions. The likelihood-ratio test allows us to compare the likelihood of the empirical data fitting two competing distributions. That is, it establishes which distribution is more likely to fit the data, and whether the difference is significant.

Following the approach described in [23], our study compares the likelihood of 5 different skewed distributions. Our hypothesis is that the power law is too "ambitious" for the observations of the tail. We also expect the distribution to be heavy tailed, i.e. with a decrease of the tail slower than in an exponential distribution. In addition to these two distributions that frame the expected tail of our data, our study adds three ~~potential~~ skewed distributions that would lie in between, presenting a slower decrease in the tail than the exponential but a stronger decrease than the power law: thetruncated power law (also named power law with exponential cut-off), the log-normal and the stretched exponential. Both the truncated power law and the log-normal distributions have two terms, while the power law, exponential and stretched exponential have only one. The number of terms of the distributions is relevant, since it is a factor for fitness.

The study exhaustively compares, for each wiki, the fit of the data to those five skewed distributions (power law, truncated power law, log-normal, exponential and stretched exponential), and identifies when the likelihood differences are statistically significant. It uses Vuong [27] method, which considers the variance of the data, and returns a p-value that states if the likelihood differences may be due to the data fluctuations, or are significant in order to favor one distribution over the other[3]. As in [23], we consider significant the differences with a p-value smaller than 0.1, i.e. those that have less than 10% probabilities of being a result of the data fluctuations. Additionally, in order to avoid over-fitting to the tail of the distribution, we force the method to fit every contributor with at least 10 contributions.

This study was performed using the Powerlaw python package [28]. Similar to the previous subsection, the python script source code, required for the performed analysis, is available as free/open source software to facilitate replication[4].

### B. DATA COLLECTION

This work investigates the distribution of participation in wikis from Wikia studying the number of edits per user. Wikia is a suitable research object to draw conclusions about participation in wikis in general. As argued by Shaw and Mako Hill [1], Wikia is an ideal setting in which to study peer production. Wikia only hosts publicly accessible, openly-licensed, volunteer-produced, peer production projects. To date, it is the largest and more diverse repository of open knowledge peer production, with a rich ecosystem of a broad diversity of topics, languages, community and wiki sizes. Furthermore, Wikia never restricts viewership, nor participation (except that from spammers or vandals). Wikia hosts some of the largest and most successful wikis in multiple topics and languages, such as Marvel or Star Wars fandom wikis, LyricWiki on song lyrics, Proteins scientific wiki, or AmericanFootballDatabase on such sport.

To collect our data we used the publicly available Wikia census described in [29] and retrieved on the 20th of February 2018[5]. However, as explained in Section II-A1, we limit our analysis to wikis with at least 100 registered users which have done at least one edit, and excluding bot users.

Thus, starting from this census data, and complementing it with additional information as explained below, we have created a new dataset to study the distribution of participation, i.e. which is the distribution of edits made by registered users, excluding bots. This dataset is complete, since it includes all the Wikia wikis with at least 100 users which made at least one contribution, resulting in ~6,000 wikis, as explained in detail below.

The mentioned Wikia census provides information of around 300,000 wikis. However, the census does not provide information on the number of edits of each user in each wiki. Thus, such information needs to be generated manually to complement the dataset. Therefore, in order to retrieve the required data, we need to query the API of each of the wikis hosted in Wikia. Spefically, we need to query the Special:ListUsers API endpoint that every MediaWiki wiki has[6]. Such Special:ListUsers page lists the information of every registered user in a given wiki, e.g. username, number of edits, groups she belongs to, or date of last edit made. A perl script was developed in order to use that endpoint and

[2]Goodness of fit tests script: ANONYMIZED

[3]The method is adapted in Clauset et al's for nested distributions such as power law and truncated power law, where a family of distributions is a subset of the other. This adapted method allows to state whether the larger family is indeed needed or both distributions are good models.

[4]Likelihood-ratio test script: ANONYMIZED

[5]Wikia census: https://www.kaggle.com/abeserra/wikia-census

[6]Note all Wikia wikis use the same wiki software, MediaWiki, maintained by Wikimedia Foundation and used by its projects, including Wikipedia.

obtain the number of edits performed by each registered user. In particular, the script queries the endpoint making a request for all users. Afterwards, it filters out the bot users, removing the users belonging to the *bot* and *bot-global* groups. As with the previous scripts, this perl script source code is available as free/open source software to facilitate replication[7]

.

The data collection was performed the 6th of November 2018 and its result is available at [8]. It contains information about $295,658$ wikis, as $8,433$ wikis endpoints were unavailable.

This data, the census wikis with the edits information, was curated to avoid duplicates and to filter wikis without human participation. After removing wikis without human participation, without statistical information and duplicates, the collection contains information about $282,039$ wikis.

Reliability of the data collected is considered high. Edit numbers are as reliable as Wikia publicly accessible statistics through the Special:ListUsers endpoint are. We have also done a consistent effort in bot identification in order to filter them out.

For statistical reasons already explained in Section II-A1, this work considers only wikis with at least $100$ registered (non-bot) users. Thus, the number of considered wikis was further reduced to $6,676$. It is important to remark that this is not a sample, but the observed population of wikis with at least 100 registered users with contributions in Wikia.

## III. RESULTS OF THE STATISTICAL TESTS

According to the goodness of fit test described in Section II-A1, the power law is a plausible distribution (i.e. it cannot be ruled out) for the 83% of the 6,676 Wikia wikis with at least 100 registered non-bot users. However, as explained in Section II-A2, that does not mean that the power law is the best choice, since other distributions may fit better the empirical data.

Thus, we perform the likelihood-ratio test to compare the pairs of the five candidate distributions as explained in Section II-A2. The distributions are power law, truncated power law, exponential, stretched exponential and log-normal. For each wiki, we perform likelihood-ratio tests comparing all the competing distributions against each other, that is, we perform 10 likelihood-ratio tests for each wiki, since there are 10 possible couples.

Figure 1 summarizes the results of these comparisons. The figure's pentagon apexes shows each of the five considered distributions. An arrow from distribution A to distribution B represents the percentage of wikis in which distribution A was preferred over distribution B in the likelihood-ratio test, while the opposite arrow represents the percentage of wikis where distribution B was superior~~significantly better~~ than distribution A. Note in some cases, the likelihood-ratio test may be inconclusive to determine which of the two

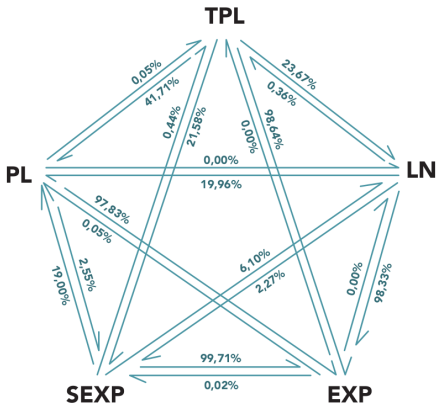


FIGURE 1: Results of the likelihood-ratio test between the five considered distributions for registered users (power law: PL, truncated power law: TPL, log-normal: LN, exponential: EXP and stretched exponential: SEXP). Each arrow from A to B has the percentage of cases in which A was superior than B.

distributions is better for a given wiki, and in those cases neither A nor B is superior. It is important to remark that the test being inconclusive means that both distributions fare similarly, which could mean that both are adequate or even that both are not adequate. For the sake of clarity, the figure omits the complementary percentage where the likelihood-ratio test was inconclusive, although it can be easily calculated[9].

The analysis of the figure results shows that the power law is not a strong contender, as it is rarely a more likely distribution than any of its competitors, with the exception of the exponential distribution, which is also overwhelmingly defeated by the rest of the candidates.

The defeat of the exponential distribution by all candidates means that a large tail of core users is clearly present in the wiki participation distributions, and that an exponential distribution, which is not able to represent heavy tails, is not a good candidate.

However, the power law being defeated by the rest of heavy-tailed distributions means that the tail is not as heavy or large as a power law would predict. Hence, more moderated heavy-tailed distributions are needed. This conclusion is similar to the one drawn in recent works that disprove the supposed prevalence of the power law in other domains [23], [24].

Thus, a correct characterization of the distributions, in nearly all cases, lies in between the exponential and the power law distributions. Among the rest of the candidates, the truncated power law stands out, since as seen in Figure 1, it is rarely beaten by its competitors: $2.16\%$ against the stretched exponential, $2.08\%$ against the log-normal, $0.18\%$ against the

[7]Script to retrieve user contributions : ANONYMIZED
[8]https://www.kaggle.com/atenorio/wikia-participation-data-20181106

[9]In all cases, percentage of A>B + percentage of A<B + percentage of inconclusive = 100%

| Distribution | Win all tests | Lose at least one test |
|---|---|---|
| Power law | 0 (0%) | 2816 (42,18%) |
| Truncated power law | 596 (8.93%) | 177 (2,65%) |
| Log-normal | 41 (0.61%) | 1159 (17.36%) |
| Stretched exponential | 2 (0.03%) | 1492 (22,35%) |
| Exponential | 0 (0%) | 6578 (98.53%) |

TABLE 1: Aggregated results of the likelihood-ratio tests for each wiki counting the cases where a candidate distribution wins all tests and loses at least one test

exponential, and $0.04\%$ against the power law distribution. Hence, the likelihood-ratio test clearly supports the truncated power law as the most appropriate distribution to characterize participation.

The appropriateness of the truncated power law is better appreciated when we aggregate the results of the likelihood-ratio tests for each wiki as shown in Table 1. We count the cases where a candidate distribution won all the likelihood-ratio tests for each wiki, which means that that distribution is the right choice for that wiki. In addition, we also counted the times where a candidate distribution lost at least one test, which means that for that wiki the candidate distribution was not the best choice.

It is important to remark that only in 10 wikis (0.15%) no candidate distribution won any likelihood-ratio test which means that they all were equally good (or, more precisely, bad) candidates. We have inspected these cases and they all exhibit uncommon participation distributions.

According to Table 1, the truncated power law is significantly better than all the candidates in 596 wikis out of the 6,676, i.e. approx. $9\%$ of the wikis considered. While the rest of the distributions fare much worse: only the log-normal and stretched exponential distributions are the best candidates in 41 and 2 wikis, respectively. The power law and the exponential are not the best candidates for any wiki, which reinforces the idea of the suitability of a heavy-tailed distribution but not as heavy as that from the power law.

According to the aggregated results in Table 1, the truncated power law is not the best or among the best candidates for only 177 wikis out of $6,676$ wikis ($2.65\%$); more precisely in 67 wikis ($1\%$) looses one test, in 101 ($1.51\%$) wikis loses two tests and in 9 ($0.1\%$)wikis looses three tests. The rest of the distributions fare much worst, e.g. log-normal can be ruled out as the best candidate in the $17.36\%$ of the wikis and the stretched exponential in the $22.73\%$. This result reinforces the idea of the truncated power law being the distribution of choice when trying to characterize the participation distribution in wikis, because it seems difficult to find a better one for most of the cases.

We show an example of participation distribution where the truncated power law won all the tests in Figure 2. The figure shows a log-log plot of the complementary distribution function where the X axis represents the number of edits in the wiki in logarithms and the Y axis the inverse cumulative relative frequency (that is, the percentage of contributors that made at least X edits in the wiki). The figure displays the

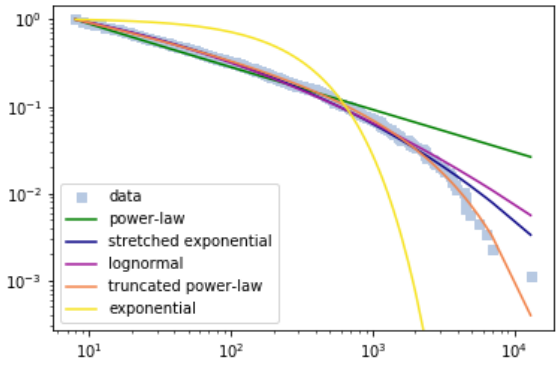


FIGURE 2: Complementary cumulative distribution function of participation of a wiki and the fitted distributions.

observations (grey squares) and the fitted distributions, the truncated power law and all the candidate distributions. The observations in the left represent the contributors with fewer edits, while those in the right are the core contributors that made more edits, i.e., the tail of the participation distribution.

It can be observed how the contributors at some point do not follow the initial slope that represents the participation distribution of the occasional contributors. While most distributions fit the initial slope, only the truncated power law is able to grasp the tail behavior. The rest of the heavy-tailed distributions predict a heavier tail, while the exponential with his bounded tail is not able to fit the community behavior at all.

While the participation distribution in Figure 2 is one of the $9\%$ examples where the truncated power law is the most convenient distributions. In most of the cases $(97,35\%)$, the Truncated Power law is not defeated by any other distribution. So there is not stastistical evidence to reject their consideration. These cases typically correspond with participation distribution with tails that can be conveniently fitted by the truncated power law, but also by the log-normal and/or the stretched exponential. The statistical analysis carried out evidences that the truncated power law is the best distribution to characterize the participation in wikis among those considered. In the next section, we will interpret the parameters of this distribution in the context of participation and will relate them with features of the wiki project.

## IV. ANALYSIS OF THE TRUNCATED POWER LAW FOR CHARACTERIZING PARTICIPATION DISTRIBUTIONS

### A. INTERPRETATION OF THE TRUNCATED POWER LAW PARAMETERS

A truncated power law is defined as a power law multiplied by an exponential: $x^{-\alpha}e^{-\lambda x}$. In the log-log plot, the parameter $\alpha$ is related to the slope of the power law function, while the parameter $\lambda$ is related to the  of the decay in the tail.

As a result, lower alphas can be associated with a more numerous cohort of occasional contributors, as their frequency

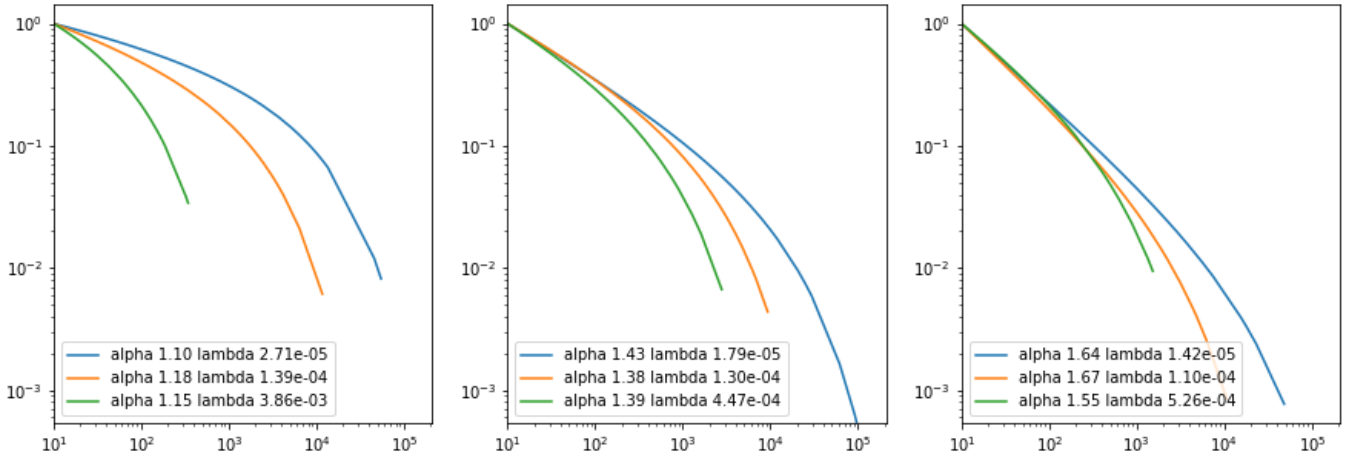


FIGURE 3: Complementary cumulative distribution functions in logarithmic scales of truncated power laws. Each sub-figure plots three wikis with similar $\alpha$ parameter, adopting smaller values in the left plot, average values in the middle and higher values in the right.

decreases less conspicuously as the number of contributions increase than in the case of higher alphas. In other words, in communities with lower alphas the frequency of contributors with more contributions decreases less markedly.

On the other hand, higher lambdas can be associated with more pronounced deviations from the power law in the tail, which means that more active contributors are less frequent as what the power law would predict. Thus, higher lambdas relate to a more numerous elite of very active contributors.

In Figure 3, we show the truncated power law of nine wikis with different $\alpha$ and $\lambda$ parameters that illustrate how diverse can be the participation distributions in wikis. From left to right we show three plots each of them with three participation distributions with roughly similar $\alpha$ values (the alpha values grow from the left to the right plot). In each plot, we show participation distributions with similar $\alpha$ but with different $\lambda$ values. This figure illustrates the idea that the initial slope of the distributions depends on $\alpha$ values, as it is steeper from the left to the right plots. While in each figure we can appreciate that higher values in the $\lambda$ parameter are associated with a more pronounced and earlier decay sooner, or, conversely, smaller values allow the power law relationship to prevail longer.

### *B. RELATIONSHIPS OF THE PARAMETERS WITH SOME FEATURES OF THE WIKI PROJECT*

Now we explore whether the $\alpha$ and $\lambda$ parameters are related to some features of the wiki project, namely, the number of edits and the number of participants. We will use scatter plots where each dot represents a wiki in a 2-dimensional plot where the axes represent the values of the $\alpha$ and $\lambda$ parameters and the dot is colored according to a color gradient related with the wiki feature. More precisely, in Figure 4 the color represents the number of edits, and in Figure 5, it represents the number of users of the wiki. For the sake of clarity, the

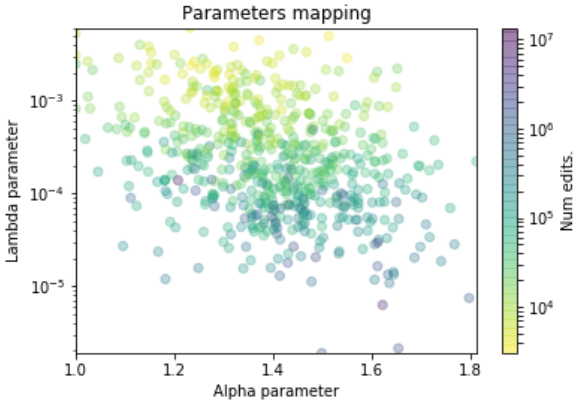


FIGURE 4: Scatter plot of the TPL-distributed wikis where the color represents the number of edits.

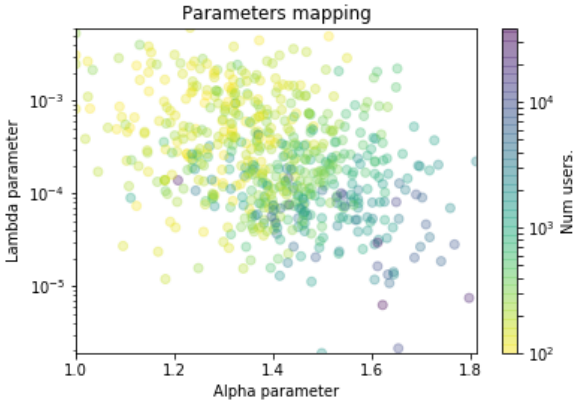


FIGURE 5: Scatter plot of the TPL-distributed wikis where the color represents the number of contributors.

plot will only display the wikis where the truncated power law distribution won all the likelihood-ratio tests.

The scatter plots show a cloud of dots with no clear relationship among the parameters. The relationship could be inverse, since wikis with large $\alpha$ and $\lambda$ values or with $\alpha$ and $\lambda$ values are rare. However, since variability is very high the inverse relationship cannot be statistically inferred.

When studying the relationship of the parameters with the size of the community in Figure 4, we can observe how the $\lambda$ parameter seems to be inversely related to the number of edits of the wiki, as the biggest wikis are distributed in the lower part of the figure and vice versa. In other words, bigger wikis (those with millions of edits) have smaller lambdas, which means that the decay in the tail of their participation distributions is not as marked. It reveals that, given an alpha value, there are more core contributors than in wikis whose participation distributions have higher lambda values, and that results in more productive communities in terms of edits. On the contrary, wikis with bigger lambdas have a less populated elite of core contributors which results in smaller wikis in terms of edits.

At Figure 5, we can perceive that the number of users of the wiki is related to the combination of both parameters, as we can see that the color gradient evolves from the upper-left towards the bottom-right corner. Participation distributions characterized by high alpha values and low lambda values belong mostly to bigger wiki communities (blue dots). Such parameter values determine an extremely sharp decrease in the (relative) frequency of editors as the number of edits increases and also a more pronounced decay on the frequency of the most active contributors. In other words, extremely unequal participation distributions can be found mostly in big wiki communities. Conversely, we can find that less unequal distributions of participation (those with low alpha and high lambda values) characterize mostly the distribution of participation of wikis with smaller communities (yellow dots).

We cannot conclude if higher unequality is cause or consequence of bigger communities and vice versa. This would require further research. However, it seems that there is a clear link between community size and participation distribution.

Furthermore, it is important to bear in mind that we are observing the participation distribution during the whole life of the wiki, that is, the aggregated effect of different communities that interacted in the wiki along time. Bigger communities are usually older communities. In this sense, it would be interesting to observe how the yearly participation distribution in these wikis evolved, because the marked inequality could be the result of the aggregation through years of more egalitarian distributions of participation.

## V. CONCLUDING REMARKS

In this work, we have critically studied the distribution of participation in wikis. We have used an extensive and diverse population of 6,676 wikis from Wikia to perform our statistical analysis. Our analysis have followed the approach defined by Clauset et al. [23]. According to the results, the power law is not an appropriate distribution, as it predicts a higher proportion of most active users than the observed in these communities. This contraditcs the bulk of the peer production literature.

From the considered distributions, the truncated power law is clearly the best according to the statistical analysis. Consequently, it should be considered as the distribution of participation of choice when characterizing wiki communities. Obviously, it may be not adequate for some specific communities, but it has been able to characterize effectively a vast majority of them and the other candidates performed significantly worse. In our analysis, we have found that the parameters of the truncated power law distribution (that govern the slope and the decay of the power law relationship in a wiki project) are related with the number of members in the community and the number of edits in the project. However, the reasons behind these findings are amatter of future research.

The prevalence of the truncated power law as the distribution of choice for characterizing the participation distribution in wikis has several implications:

- The truncated power law implies that the power law behavior holds true only in a limited range and that from that point a decay can be observed. In a distribution of participation, it means that the truncated power law fits better not only the frequency of the occasional contributors of the community, but also the frequency of the most active ones. The change of slope may also serve to empirically determine a division between core and non-core contributors instead of using arbitrary divisions e.g. as in other studies [2]. Further research may provide insights on why the inner dynamics change, and how we can study better the different emergent roles within peer production communities.
- In a truncated power law, core contributors are rarer than would be in a power law with the same slope. It means that in the tail the decrease in the frequency of contributors as the edit activity increases is more marked. It seems to reinforce the idea that core contributors are special, in the sense, that they are very few and may have different motivations than those had by the rest of the community. The reasons behind could be due to community dynamics such as some kind of elitism that prevents more people to be involved as much as those more active in the community, or that the many active users experiment a burnout at some point and cease or decrease their activity level.

The approach followed by this work has several limitations:

- It is a descriptive quantitative work, and thus it lacks explanatory aspects that further qualitative research could contribute with.
- We are cautious with the generalizability of our findings, and yet, considering the significant size and diversity of

the sample used, and similar generalizations performed in the field, e.g. [1], we may state that these results need to be validated in other other parts of the wikisphere, such as the Wikimedia Foundation projects, which may exhibit a different participation distribution, as well as in other peer production communities such as Free Software projects. Thus, we encourage other researchers to replicate our approach with other peer production communities.

- The statistical analysis methods employed require a certain wiki size to have conclusive results, which may constrain their applicability for smaller wikis. Despite of having 300,000 wikis in Wikia, most of them are under 100 users and thus are discarded, using "only" 6,676 wikis in the analysis.
- We have analyzed the participation in the communities aggregated through time (years), that is, accumulating the participation of all the members from the beginning. However, the members of a wiki community change through time, as change the participation dynamics. The participation distribution could be very different when analyzed in a smaller time window, for example, a year.

We have already defined several potential lines for future work, but we would like to mention those that we consider more interesting:

- Use a different base population, in order to appropriately generalize for peer production communities and not just wikis. For instance, we could analyze in a similar manner communities from Github, Wikimedia Foundations projects, or Stack Exchange.
- Perform a temporal analysis with a rolling time window, to understand how these distributions evolve over time, especially considering the evolution of the truncated power law parameters and how they relate with participation dynamics and inequality.
- Study the characterization of wikis based on their truncated power law parameters, clustering similar wikis and explaining the causes or consequences of the different typologies and how they relate with factors such as maturity stage, community dynamics and sustainability. Such analysis are not

We can conclude the truncated power law is more appropriate than any other distribution to represent the distribution of participation in wikis from Wikia. Our results can be better understood if they are put in context with a previous study that questioned the prevalence power law in several fields [23] and the ground-breaking finding that the power-law was indeed rare in real-life networks [24]. Our study proposes the characterization of participation in wikis as a truncated power law as well as the use of rigorous tools to further validate this characterization and the need to investigate the causes behind such distribution of participation.

## REFERENCES

[1] A. Shaw and B. M. Hill, "Laboratories of oligarchy? how the iron law extends to peer production," Journal of Communication, vol. 64, no. 2, pp. 215–238, 2014.

[2] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie," World wide web, vol. 1, no. 2, p. 19, 2007.

[3] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl, "Creating, destroying, and restoring value in wikipedia," in Proceedings of the 2007 international ACM conference on Supporting group work. ACM, 2007, pp. 259–268.

[4] M. Fuster Morell, "Participation in online creation communities: Ecosystemic participation," in Conference Proceedings of JITP 2010: The Politics of Open Source, vol. 1, 2010, pp. 270–295.

[5] F. Ortega, J. M. Gonzalez-Barahona, and G. Robles, "On the inequality of contributions to wikipedia," in Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), Jan 2008, pp. 304–304.

[6] P. Neis and D. Zielstra, "Recent developments and future trends in volunteered geographic information research: The case of openstreetmap," Future Internet, vol. 6, no. 1, pp. 76–106, 2014.

[7] B. M. Hill and A. Shaw, "The wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation," PloS one, vol. 8, no. 6, p. e65782, 2013.

[8] J. Reagle, ""free as in sexist?" free culture and the gender gap," first monday, vol. 18, no. 1, 2012.

[9] O. Arazy, F. Ortega, O. Nov, L. Yeo, and A. Balila, "Functional roles and career paths in wikipedia," in Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, 2015, pp. 1092–1105.

[10] J. Preece and B. Shneiderman, "The reader-to-leader framework: Motivating technology-mediated social participation," AIS transactions on human-computer interaction, vol. 1, no. 1, p. 5, 2009.

[11] B. Vasilescu, A. Serebrenik, P. Devanbu, and V. Filkov, "How social q&a sites are changing knowledge sharing in open source software communities," in Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014, pp. 342–354.

[12] A. Serrano, J. Arroyo, and S. Hassan, "Webtool for the analysis and visualization of the evolution of wiki online communities," in Proceedings of the European Conference on Information Systems (ECIS) 2018. AIS Electronic Library (AISeL), 2018.

[13] J. Stuckman and J. Purtilo, "Analyzing the wikisphere: Methodology and data to support quantitative wiki research," Journal of the American Society for Information Science and Technology, vol. 62, no. 8, pp. 1564–1576, 2011.

[14] K. Healy and A. Schussman, "The ecology of open-source software development," Technical report, University of Arizona, USA, Tech. Rep., 2003.

[15] S. K. Sowe, I. Stamelos, and L. Angelis, "Understanding knowledge sharing activities in free/open source software projects: An empirical study," Journal of Systems and Software, vol. 81, no. 3, pp. 431–446, 2008.

[16] C. M. Schweik and R. C. English, Internet success: a study of open-source software commons. MIT Press, 2012.

[17] V. Cosentino, J. L. C. Izquierdo, and J. Cabot, "A systematic mapping study of software development with github," IEEE Access, vol. 5, pp. 7173–7192, 2017.

[18] F. Wu, D. M. Wilkinson, and B. A. Huberman, "Feedback loops of attention in peer production," in Computational Science and Engineering, 2009. CSE'09. International Conference on, vol. 4. IEEE, 2009, pp. 409–415.

[19] D. M. Wilkinson, "Strong regularities in online peer production," in Proceedings of the 9th ACM conference on Electronic commerce. ACM, 2008, pp. 302–309.

[20] J. Howison, K. Inoue, and K. Crowston, "Social dynamics of free and open source team communications," in Open Source Systems, ser. IFIP International Federation for Information Processing, E. Damiani, B. Fitzgerald, W. Scacchi, M. Scotto, and G. Succi, Eds. Springer US, Jun. 2006, no. 203, pp. 319–330. [Online]. Available: http://link.springer.com/chapter/10.1007/0-387-34226-5_32

[21] K. Crowston, K. Wei, Q. Li, and J. Howison, "Core and Periphery in Free/Libre and Open Source Software Team Communications," in Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006. HICSS '06, vol. 6, Jan. 2006, pp. 118a–118a.

[22] P. Barbrook-Johnson and A. Tenorio-Fornés, "Modelling commons-based peer production: The commoners framework," in Social Simulation Conference 2017 (SSC2017). Dublin, Ireland. European Social Simulation Association (ESSA), 2017.

[23] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," SIAM review, vol. 51, no. 4, pp. 661–703, 2009. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/070710111
[24] A. D. Broido and A. Clauset, "Scale-free networks are rare," arXiv preprint arXiv:1801.03400, 2018.
[25] F. Ortega, "Wikipedia: A quantitative analysis," Ph.D. dissertation, PhD thesis. Universidad Rey Juan Carlos, Madrid, 2009.
[26] C. S. Gillespie, "Fitting heavy tailed distributions: the powerlaw package," arXiv preprint arXiv:1407.3492, 2014.
[27] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," Econometrica: Journal of the Econometric Society, pp. 307–333, 1989.
[28] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: a python package for analysis of heavy-tailed distributions," PloS one, vol. 9, no. 1, p. e85777, 2014.
[29] G. Jiménez-Díaz, A. Serrano, and J. Arroyo, "A wikia census: motives, tools and insights," in Proceedings of Opensym 2018. ACM, 2018.

JAVIER ARROYO is Associate Professor at the Department of Software Engineering and Artificial Intelligence of the Universidad Complutense of Madrid (UCM) since 2013 and researcher in the Instituto de Tecnología del Conocimiento. He got a PhD degree in Computer Science from Universidad Pontificia Comillas (2008).

His research interests include among others time series forecasting and machine learning applied to different domains and real-life problems.

JAVIER ARROYO is Associate Professor at the Department of Software Engineering and Artificial Intelligence of the Universidad Complutense of Madrid (UCM) since 2013 and researcher in the Instituto de Tecnología del Conocimiento. He got a PhD degree in Computer Science from Universidad Pontificia Comillas (2008).

His research interests include among others time series forecasting and machine learning applied to different domains and real-life problems.